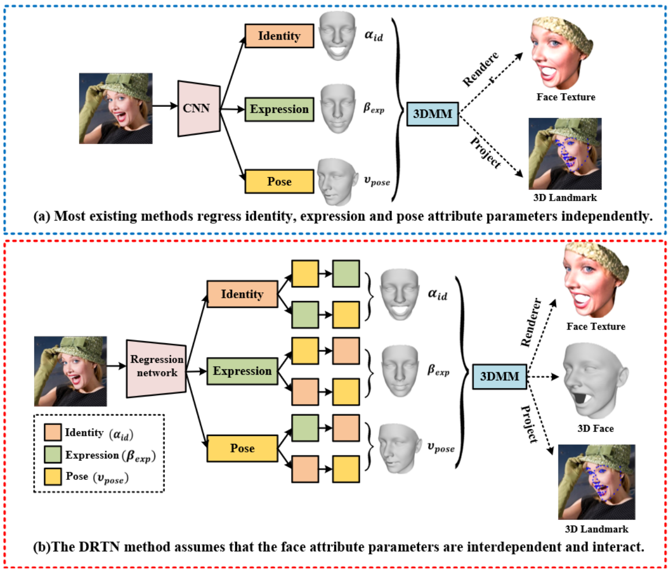
Disentangled Representation Transformer Network for 3D Face Reconstruction and Robust Dense Alignment

Xiangzheng Li, Xiaofei Li, Jia Deng, Xiangzheng Li*
School of Mathematics and Computer Science, Ningxia Normal University, Guyuan,756000, China
*Corresponding author Email: 82021011@nxnu.edu.cn

***Abstract*—In this paper, we propose a disentangled representation transformer network (DRTN) approach for 3D dense face alignment and reconstruction.Unlike traditional 3DMM-based approaches in which the target parameters, namely the shape, expression, and pose parameters, are all individually estimated, without considering their direct influences on one another, although they are jointly optimized our DRTN aims to enhance the representation of facial attributes in a semantic sense by learning the correlation of different 3D facial attribute parameters.To achieve this we present a novel strategy to design disentangled 3D face attribute representation,which decomposes the given facial attributes into identity, expression, and poses parts. Specifically, the estimate of 3D face parameters in the regression network depends on the correlation of other face attribute parameters rather than being independent. The branching of the identity component aims to reinforce the learning of expression and pose attributes by preserving the overall face geometry structure and keeping the identity intact. Accordingly, the expression and pose parts of the branch maintain the consistency of expression and pose attributes, respectively. It helps refine the reconstruction and alignment of face details in large poses mainly by coupling other facial attribute parameters. Extensive qualitative and quantitative experimental results on widely-evaluated benchmarking datasets demonstrate that our approach achieves competitive performance compared to state-of-the-art methods.**

***Index Terms*—Face alignment and reconstruction, convolutional neural networks, disentangled representation learning, features interaction, biometrics.**



(a) Most existing methods regress identity, expression and pose attribute parameters independently.

(b)The DRTN method assumes that the face attribute parameters are interdependent and interact.

Fig. 1. Conventional 3D face reconstruction and our DRTN framework. Inside the blue dashed box is the regression of face attribute parameters independently using the traditional method. In the red dashed box is the DRTN method using a correlation between face attribute features to reinforce the learning of attribute parameters by the network.

I. Introduction

3D face reconstruction is an essential topic in computer vision and graphics. The human face is one of the most discriminative parts of the human body, and the facial features and contours contain many attributes and semantic information. The single-view 3D face reconstruction targets recovering a whole face geometry shape based on the given single-view image, which plays a significant role in many visual analysis applications, such as face recognition [1],face verification [2], face expression analysis [3], and facial animation [4].

However, many face images are affected by the shooting angle and the surrounding environment. Problems such as partial occlusion of faces and facial blurring occur, resulting in distortion of reconstructed 3D faces. Therefore, how to perform accurate 3D face reconstruction and facial landmark alignment under large face poses, partial occlusion, and insufficient illumination is a problem that still needs to be solved. Conventional methods are mainly based on optimization algorithms, e.g., iterative closest point [5], Shape from Shading [6], [7],and Photometric Stereo [8]. Nevertheless, the problems of locally optimal solutions, short model initialization, and high optimization complexity of these face reconstruction techniques make the 3D face reconstruction process complex and costly.

With the rise of deep learning in recent years, CNN network regression-based methods [9], [10] have emerged to achieve remarkable success in face 3D reconstruction and dense alignment.t is very challenging to reconstruct 3D face shapes from 2D images without prior knowledge. This is mainly because 2D data does not convey clear depth information.A common approach to solving the single-view 3D face reconstruction

* represents corresponding author. This work was supported in part by the National Science Foundation of China under Grant 61662059, and in part by the National Science Foundation of China under Grant 62062056.

problem is using a set of 3D base shapes to capture the subspace or constructing a morphological model of face shape variation.Feng et al. [11] proposed a simple convolutional neural network for 3D face reconstruction and dense alignment that regresses the UV position map from a single 2D image and records the complete 3D face shape in UV space. The method does not rely on any a priori face model and can reconstruct the complete face geometry based on the semantic information of the face.Jackson et al. [12] pointed out that existing 3D face reconstruction methods are affected by the pose, expression, and illumination of the input face photos and are complex and inefficient in fitting the face model. So the method bypasses the construction of a 3D deformation model and regresses the 3D facial geometry using a single 2D facial image.However, these methods cannot explicitly capture information about individual attributes of the face, which is useful for single-view-based 3D face reconstruction and face alignment. To address this challenge, [13], [23] proposed an encoder-decoder network to separate the identity features and expression features in the 3D face reconstruction process from a single 2D image and encode them nonlinearly to accomplish accurate 3D face reconstruction.While encouraging performance has been obtained, these methods are only single regression parameters that cannot explicitly exploit the complementary information between face attributes.

In this paper, we propose a disentangled representation learning method for single-view 3D reconstruction and face alignment. Some current approaches [14] are mainly decomposing the face attributes and individually estimating their shape, expression, and pose parameters. Although this enhances the learning of a single face attribute, these methods do not consider the interaction between attributes. Motivated by extracting shape, expression, and pose parameters from 2D images and fusing the facial parameters to capture the dependencies between attributes, which is useful for 3D face reconstruction and alignment in environments such as large pose, self-obscuration, and poor lighting. Our approach aims to achieve complementary feature information by decomposing the face attribute information to enhance the correlation between individual attributes. To achieve this, we carefully design disentangled representation transformer network (DRTN), which includes identity, expression, and pose branches. The branching of the identity component aims to enhance the learning of expression and pose attributes by preserving the overall face geometry structure and keeping the identity intact. Accordingly, the expression and pose parts of the branch maintain the consistency of expression and pose attributes, respectively. It helps refine the reconstruction and alignment of face details in large poses mainly by coupling other facial attribute parameters. The network parameters of the designed attribute branch architecture are optimized by backpropagation in an end-to-end manner. **Fig. 1** shows the pipeline of our proposed DRTN. Experimental results show that the method significantly outperforms traditional independent regression of attribute parameters in 3D face reconstruction and landmark alignment and exhibits very competitive performance on the test dataset. The contributions of our work are summarized as follows:

(1) We develop a decomposed representation learning method for faces to explicitly model the correlation between face attributes. The method reduces the ambiguity of face attribute learning from traditional CNN-based parameter regression. It enhances the learning of face attribute information enabling semantic editing of identity, expression, and pose domains.

(2) We propose a novel disentangled representation transformer network framework based on single-view 3D face reconstruction and alignment. Our framework uses face attribute branching regression for the representation, independently regressed from the identity, expression, and pose components. Based on identity attribute consistency, the identity branch introduces pose and expression attribute information to enhance the integrity of the geometric face profile. Accordingly, the branches of the expression and pose parts refine the expression and landmark alignment effects of the 3D face by coupling other attribute parameters while maintaining the consistency of each attribute.

(3) To further improve the performance, our transformer network is designed to address the problem of missing details of face geometric contours. The depth model's capacity is effectively controlled, and complementary information is extracted from the multi-attribute face input. This aims to establish similarities between shallow and deep representations of faces and to mine local attribute information of faces using the implementation of global information interactions.

II. Related Works

This section briefly reviews the existing 3D facial alignment methods and reconstruction techniques.

A. *Face Alignment*

Face alignment in computer vision is a long-standing and widely discussed problem. Initially, some 2D face alignment methods, which fit the face shape of a given input image by constructing an overall fitting template, aimed to locate a set of baseline 2D facial landmarks. Representative methods of this type include active shape models (ASM) [15], active appearance models (AAM) [16],and constrained local models (CLM) [17]. However, when the pose is large, the computation is usually slow and limited in describing the face shape due to multiple classifier regressions to locate the landmarks of 2D faces.

Recently, work has used deep learning to study 3D face alignment in large poses. Among the methods for 3D face alignment is the direct regression of face parameters bypassing the 3D deformation model and fitting a dense 3DMM with CNN cascade regression. For example, Amin Jourabloo et al. [18] designed a cascaded coupled regression method to estimate the camera projection matrix and 3D face landmarks by integrating a 3D point distribution model. Adrian Bulat et al. [19] proposed a heat map regression-based method to estimate 3D face landmarks. Each landmark corresponds to a

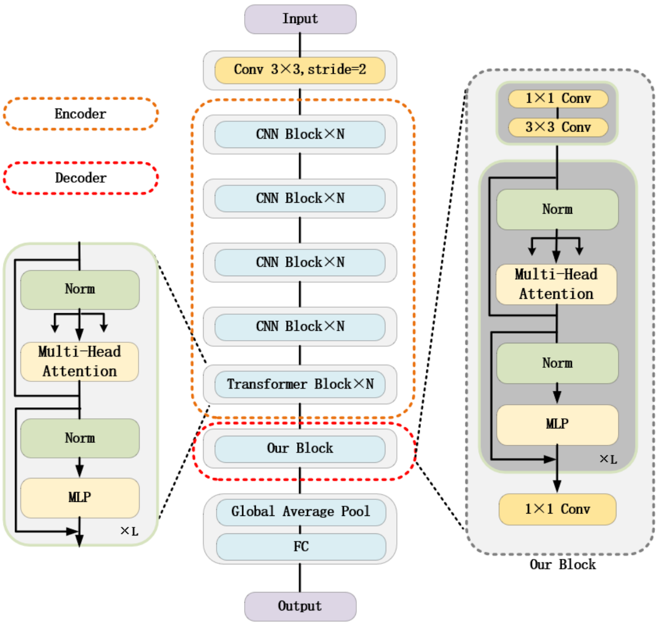


Fig. 2. Model overview. Our transformer module consists of an encoder and a decoder. In the encoder, each attribute information of the face is extracted from the face image by a multi-layer convolutional neural network. This attribute information is then passed through the transformer block to achieve uniform encoding of attribute information. Then, for the encoded face feature sequence, we will use the decoder module to extract each face attribute information in a multi-head attention mechanism and output the face attribute information through the fully connected layers.

heat map in this method, and these heat maps are regressed with the input RGB images by learning the 3D face depth values through a residual network. A video-based 3D cascade regression method has been developed by Jeni et al [20]. A dense 3D shape is generated in real-time from an input 2D face image. The algorithm estimates the position of the dense set of markers and their visibility and then achieves dense 3D face alignment by fitting a partial 3D model. Xin Ning et al. [21] proposed a real-time 3D face alignment method that uses an encoder-decoder network with efficient convolutional layers to enhance information transfer between different resolutions in the encoding and decoding stages and achieve advanced performance.

At present, several works have performed face alignment by fitting 3D variable shape models (3DMM) to 2D facial images. [22] fits a 3DMM by a single CNN in an iterative manner, while the CNN augments the input channel with the represented shape features in each iteration. [23] uses multi-constraint training CNNs to estimate 3DMM parameters and then provides very dense 3D alignment. Lei Jiang et al. [24] proposed a dual-attention mechanism and a practical end-to-end 3D face alignment framework. A stable network model is constructed by deep separable convolution, densely connected convolution, and light channel attention mechanism. This model-based 3D reconstruction method can easily accomplish the task of 3D face alignment. However, these methods only use deep networks to directly learn the relationship between 2D input images and 3D models and do not consider the integrity of the correspondence between face attributes. To address this, our model decomposes face attributes by a unified deep learning architecture and automatically extracts useful feature information directly from pixels using complementary information between face attributes. This approach inferred face contours in invisible regions and improved face alignment in large poses.

B. *3D face reconstruction*

Initially, 3D face reconstruction was mainly used in medical applications for human head diagnosis. Face reconstruction is mainly done by scanning the face with a 3D scanner [25], [26] to obtain the face's shape, structure, and texture information to reconstruct a full 3D face. Although it is more accurate to reconstruct 3D faces using scanners, the whole implementation process is more complicated and time-consuming. Therefore, Volker Blanz et al. [27] proposed the 3D morphable model (3DMM). The 3D morphable model is built based on a 3D face database, with face shape and face texture statistics as constraints. The influence of pose and illumination factors on the facial reconstruction process is considered to make the generated 3D face model more accurate. Later, Paysan et al. [28] proposed a basel face model (BFM) based on the original 3DMM with improved alignment algorithms for higher shape and texture accuracy. However, most of these methods establish the correspondence of vertices between the input image and the 3D template and then solve the nonlinear optimization function to regress the 3DMM coefficients. Therefore these methods rely heavily on the accuracy of landmarks or other feature point detectors.

Due to the wide application of deep learning in various fields, many works have recently utilized CNNs to predict face parameters for reconstruction directly. Elad et al. [29] introduced an end-to-end convolutional neural net framework, which generates the geometry of a face in a coarse to fine manner. Subsequently, Dou et al. [30] proposed a deep neural network-based approach to improve facial expression reconstruction by integrating multi-task loss functions and fused convolutional neural networks into a DNN structure. This approach avoids the complex 3D rendering process, but the reconstruction process is only valid for frontal faces. Tran et al. [31] made a nonlinear improvement to the traditional linear 3DMM model by performing end-to-end learning in a weakly supervised manner. Lee et al. [32] proposed to use an uncertain-perception encoder that effectively combines graph convolutional neural networks and generative adversarial networks. Wu et al. [33] propose a process for learning 3D deformation models from original single-view images without external supervised learning. Browatzki [34] proposes a semi-supervised approach. The key idea is to generate implicit face information from many existing unlabeled face images. Feng et al. [11] used UV maps to map 3D shapes to 2D images for representation and then constructed 3D face shapes. However, these methods do not perform well with large poses and strongly occluded faces. The Jackson et al. [12] method bypasses the problems associated with the construction and fitting of 3D deformation models by using a single 2D face image to de-regress the volume of the 3D face geometry for face reconstruction. The method is no longer restricted to the model space but requires a complex network structure and much time to predict the voxel data. Recently, some works decompose a given 3D face into identity and expression parts and encode them nonlinearly to achieve 3D face reconstruction with remarkable results. However, these methods do not consider the interaction between face attributes and only estimate the attributes individually in the process of parameter regression. Unlike the above methods, our approach enhances the semantically meaningful face attribute representation. It directly obtains the complete 3D face geometry and its corresponding information by learning the correlation of different 3D face attribute parameters.

III. Proposed Method

In this section, we first introduce a 3D face model with potential representations and design our approach based on it. Then we propose a transformer-based joint learning pipeline for encoders and decoders. Finally, specific implementation details of this face attribute branching method are given, including the network structure, training data, and training procedure.

A. A Composite 3D Face Shape Model

The 3D morphable model is one of the most successful methods for describing 3D face reconstruction. In this work, we adopt a common practice in 3D morphable model (3DMM) [27], representing a 3D human face as a combination of shapes and expressions. The concatenation of its vertex coordinates represents each 3D face shape as:

$$S=\left[x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_n, y_n, z_n\right]^T \tag{1}$$

where n is the number of vertices in the point cloud of the 3D face, and T is transpose. $S_i=(x_i, y_i, z_i)$ denotes the coordinates of (x, y, z) in the cartesian coordinate system.In this paper, we use a 3D morphable model (3DMM) [27] to recover the 3D geometry of a human face from a single image. We use the 3DMM PCA model to represent the face geometry S_{Model} as:

$$\begin{aligned} S_{\text{Model}} & =\bar{S}+\Delta S_{id}+\Delta E_{\exp } \\ & =\bar{S}+A_{id} \alpha_{id}+B_{exp} \beta_{\exp } \end{aligned} \tag{2}$$

where $\bar{S} \in R^{3n}$ is the mean geometry. $\Delta S_{id}$is the identity-sensitive difference between $\bar{S}$ and S_{Model} , and $\Delta E_{\exp}$ dnotes the expression-sensitive difference. A_{id} and α_{id} are the identity base and identity parameters of the face. B_{exp} and $\beta_{\exp}$ are the expression base and expression parameters of the face. $\bar{S}$ and A_{id} are learned from Basel Face Model [28]and B_{exp} is obtained from FaceWarehouse [36].

In the process of 3D face fitting, we use a weak perspective projection to project the 3D face onto the 2D face plane. This process is denoted as follows:

$$C_{(p)}=f * P_r * R * S+t \tag{3}$$

where C is the geometry projected in image coordinates, f is a scale factor, $P_r=\begin{pmatrix}1 & 0 & 0 \\ 0 & 1 & 1\end{pmatrix}$ is an orthogonal projection, R is a rotation matrix consisting of 9 parameters, and t is a translation vector. We can then transform the 3D face reconstruction problem into a face parameter regression problem. We have 62 parameters to regress on the 3D face model, where the pose parameter $v_{pose}=[f, R, t]$, so the set of all model parameters is $p=\left[v_{pose}, \boldsymbol{\alpha}_{id}, \boldsymbol{\beta}_{\exp}\right]^T$.

B. Transformer module

Previous approaches [23] used Disentangled representation to extract individual face attributes from face images. However, this approach cannot simultaneously model shape, expression, and texture without decomposing the underlying representation into relevant factors such as identity and expression while preserving identity information. Therefore, we use an encoder to extract three different sets of features, i.e., identity, expression, and pose. Inspired by [37], the Transformer structure can extract global information from the input features and implement information exchange within each entry. In contrast, the tiny sensory field of convolutional neural networks can only mine local information. In order to achieve an effective combination of global and local information, we introduce the Transformer module inside the encoder and decoder. This module can obtain the global representation from the shallow layer and extract the geometric structure information of the face. In addition, this module can obtain more similarity of face attributes between shallow and deep representations and extract deep face semantic information.

As shown in **Fig2**, we introduce the Transformer module inside the encoder and decoder, which plays a vital role in the 3D face space. The semantic features in the image are extracted layer by layer inside the encoder in a structured way with a multi-layer convolutional connection Transformer block. Specifically, the encoder module is divided into four layer layers and a Transformer structure. Firstly, the input face image is convolved by Conv 3×3 to obtain the initial features of the face. Then the encoder network extracts the initial face features layer by layer to gradually increase the number of face features channels. Finally, the Transformer block is mapped to the high-dimensional space, where the Transformer structure encodes the face as a series of $2048\times7\times7$ feature sequences. In addition, the traditional Transformer structure uses a self-attentive mechanism to directly compute the attention weights at each position of the sentence in the encoding process by some operations; and then compute the implied vector representation of the whole sentence in the form of a sum of weights. However, the drawback of this self-attentive mechanism is that the model overly focuses on a single attribute position of the face when encoding the information at the current position. Therefore, we use a multi-headed attention mechanism to learn the different attribute behaviors of faces and then combine the behaviors with different face attributes as knowledge. The aim is to capture the dependencies between

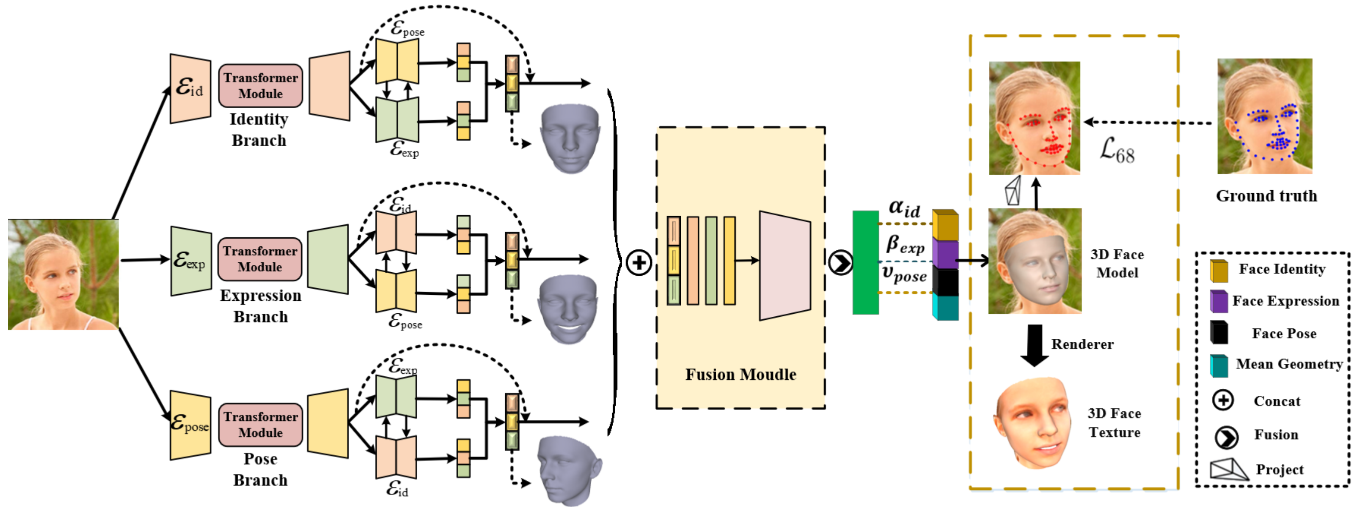


Fig. 3. The pipeline of the proposed disentangled representation transformer network (DRTN). Our network consists of two parts, the decomposition part, and the fusion part. The decomposition part is divided into three branches, one is for extracting expression and pose information based on identity, another is for extracting identity and pose features based on expression, and the last is for extracting identity and expression features based on the pose. The fusion module aims to obtain the face information related to each attribute from the output of the identity, expression, and pose branch networks and use the fusion module to merge the face attribute information to complete the 3D face reconstruction. During learning phase, we use Euclidean distance loss $\mathcal{L}_{68}$ to constrain the geometry of the face.

individual attributes of the face within a sequence and improve the subspace representation.

Besides, our module preserves the local features of the face image in order not to destroy the spatial structure of its image. In the decoder, we do not encode the image with position vectors as in the TRT-ViT [37] approach but convolve the pre-passed face feature sequence by Conv1×1 and then input the obtained feature sequence into the transformer structure. Immediately afterward, the transformer module dimensionally transforms the input face attribute features and checks the distribution of each face attribute from the feature space, then calculates the correlation between the face attribute feature points.

C. *Attribute Decomposition Representation*

Single-view 3D face reconstruction is to obtain estimates of shape α_{id}, expression β_{exp}, and pose parameters v_{pose} given an input image I. A wide variety of attribute sources in the face may lead to variability in the facial reconstruction. For example, the distortion of the geometric contours of the face that expressions may cause. Therefore, our goal is to obtain the representational values of each potential attribute from the face image by a function $\mathcal{F}_i$ given the input image $\boldsymbol{I}$ as:

$$\left[\bar{E}_{id}, \bar{E}_{exp}, \bar{E}_{pose}\right] = \mathcal{F}_i\left(\boldsymbol{I}; \boldsymbol{\alpha}_{id}, \boldsymbol{\beta}_{exp}, \boldsymbol{v}_{pose}\right) \tag{4}$$

where α_{id}, β_{exp}, and v_{pose} are the identity, expression, and pose parameters involved in $\mathcal{F}_i$. Usually, the potential face attributes represent $\bar{E}_{id}$, $\bar{E}_{exp}$, and $\bar{E}_{pose}$ in a much lower dimension than the input 2D face image $\boldsymbol{I}$ and the output 3D face shape S. Alternatively, previous approaches extracted shape, expression, and pose attribute information independently given the input image $\boldsymbol{I}$, i.e., $\left[\bar{\boldsymbol{E}}_{id}\right] = \mathcal{F}_{id}\left(\boldsymbol{I}; \boldsymbol{\alpha}_{id}\right)$, $\left[\bar{\boldsymbol{E}}_{exp}\right] = \mathcal{F}_{exp}\left(\boldsymbol{I}; \boldsymbol{\alpha}_{exp}\right)$ and $\left[\bar{\boldsymbol{E}}_{pose}\right] = \mathcal{F}_{pose}\left(\boldsymbol{I}; \boldsymbol{\alpha}_{pose}\right)$.

However, this simple feature extraction does not consider the correlation between face attributes but only decomposes the low-dimensional 3D faces between face attribute variables. We designed a disentangled representation transformer network (DRTN) to solve this problem. The network structure is shown in **Fig 3**, and the dependencies between face attributes are learned by regression from the identity, expression, and gesture branches, respectively. The specific learning of face attributes for each branch is shown as follows.

Identity branch: One of the branches of the identity component aims to enhance the learning of expression and pose attributes by preserving the overall face geometry structure and keeping the identity unchanged. In the identity branch, we explicitly model the individual face attribute dependencies. Where the joint expression and pose attributes are decomposed under the condition that the identity attribute $\bar{E}_{id} = \mathcal{F}_{id}\left(\boldsymbol{I}; \boldsymbol{\alpha}_{id}\right)$ is consistent as:

$$\bar{E}_{id,exp} = \mathcal{F}_{id,exp}\left(\boldsymbol{\beta}_{\mathrm{exp}}\, ; \boldsymbol{I}, \bar{E}_{id}\right) \tag{5}$$

$$\bar{E}_{id,exp,pose} = \mathcal{F}_{id,exp,pose}\left(v_{pose}; \boldsymbol{I}, \bar{E}_{id}, \bar{E}_{id,exp}\right) \tag{6}$$

$$\bar{E}_{id,pose} = \mathcal{F}_{id,pose}\left(v_{pose}; \boldsymbol{I}, \bar{E}_{id}\right) \tag{7}$$

$$\bar{E}_{id,pose,exp} = \mathcal{F}_{id,pose,exp}\left(\boldsymbol{\beta}_{exp}; \boldsymbol{I}, \bar{E}_{id}, \bar{E}_{id,pose}\right) \tag{8}$$

we formulate the learning process of these parameters with three autoencoders $\mathcal{E}_{id}$, $\mathcal{E}_{exp}$, and $\mathcal{E}_{pose}$. $\bar{E}_{id}$ is the identity attribute information learned from the input image $\boldsymbol{I}$ after passing through the $\mathcal{E}_{id}$. $\bar{E}_{id,exp}$ is the expression information obtained by $\mathcal{E}_{exp}$ encoder learning on the basis of the identity attribute $\bar{E}_{id}$. The $\bar{E}_{id,exp,pose}$ is the pose information learned by the $\mathcal{E}_{pose}$ encoder on the basis of $\bar{E}_{id,exp}$. Similarly, $\bar{E}_{id,pose}$ is the pose information obtained by $\mathcal{E}_{pose}$ encoder learning on the basis of the identity property $\bar{E}_{id}$. $\bar{E}_{id,exp,pose}$

is the expression information obtained by $\mathcal{E}_{exp}$ encoder learning on the basis of $\bar{E}_{id,pose}$. $\mathcal{F}_{\boldsymbol{i}}(\cdot)$ is the learnable encoder among the autoencoders that have gone through different orders.

Although this method can effectively solve the problem that information between face attributes cannot interact with each other, however, the parameter estimation using this sequential decomposition method is somewhat scattered. In this problem, we use the coupled variable method to fuse the expression and posture attribute information under the condition of identity invariance, and the fused attribute representation value as :

$$\bar{T}_{(\alpha_{id},\beta_{exp},\upsilon_{pose})} = \bar{E}_{id} \otimes \bar{E}_{id,exp} \otimes \bar{E}_{id,exp,pose} \tag{9}$$

$$\bar{T}_{(\alpha_{id},\upsilon_{pose},\beta_{exp})} = \bar{E}_{id} \otimes \bar{E}_{id,pose} \otimes \bar{E}_{id,pose,exp} \tag{10}$$

where $\bar{T}\left(\alpha_{id},\beta_{exp},\upsilon_{pose}\right)$ and $\bar{T}\left(\alpha_{id},\upsilon_{pose},\beta_{exp}\right)$ represent the face expression and pose features obtained after the facial attributes pass through different encoder networks based on identity consistency. where$\otimes$ is the element-wise Hadamard product.

Expression branch: Correspondingly, the branch of expression part couples the identity and pose attributes to refine the face facial details reconstruction while preserving the consistency of expression attribute $\bar{E}_{exp} = \mathcal{F}_{\text{exp}}\left(\boldsymbol{I};\boldsymbol{\beta}_{\text{exp}}\right)$. Its joint attribute decomposition is expressed as:

$$\bar{E}_{exp,id} = \mathcal{F}_{exp,id}\left(\alpha_{id};\boldsymbol{I},\bar{E}_{exp}\right) \tag{11}$$

$$\bar{E}_{exp,id,pose} = \mathcal{F}_{exp,id,pose}\left(\upsilon_{pose};\boldsymbol{I},\bar{E}_{exp},\bar{E}_{exp,id}\right) \tag{12}$$

$$\bar{E}_{exp,pose} = \mathcal{F}_{exp,pose}\left(\upsilon_{pose};\boldsymbol{I},\bar{E}_{exp}\right) \tag{13}$$

$$\bar{E}_{exp,pose,id} = \mathcal{F}_{exp,pose,id}\left(\alpha_{id};\boldsymbol{I},\bar{E}_{exp},\bar{E}_{exp,pose}\right) \tag{14}$$

where $\bar{E}_{exp}$ is the expression attribute information obtained from the input image $\boldsymbol{I}$ after $\mathcal{E}_{exp}$ encoder learning. $\bar{E}_{exp,id}$ is the identity information obtained by $\mathcal{E}_{id}$ encoder learning on the basis of the expression attribute $\bar{E}_{exp}$. $\bar{E}_{exp,id,pose}$ is the pose information obtained by $\mathcal{E}_{id}$ encoder learning on the basis of $\bar{E}_{exp,id,pose}$. Similarly, $\bar{E}_{exp,pose}$ is the pose information obtained by $\mathcal{E}_{pose}$ encoder learning on the basis of the expression property $\bar{E}_{exp}$. $\bar{E}_{exp,pose,id}$ is the identity information obtained by $\mathcal{E}_{id}$ encoder learning on the basis of $\bar{E}_{exp,pose}$. Simultaneously, the representational values of their identity and pose attributes under the condition of constant expression attributes are :

$$\bar{T}_{(\beta_{exp},\alpha_{id},\upsilon_{pose})} = \bar{E}_{exp} \otimes \bar{E}_{exp,id} \otimes \bar{E}_{exp,id,pose} \tag{15}$$

$$\bar{T}_{(\beta_{exp},\upsilon_{pose},\alpha_{id})} = \bar{E}_{exp} \otimes \bar{E}_{exp,pose} \otimes \bar{E}_{exp,pose,id} \tag{16}$$

where $\bar{T}\left(\beta_{exp},\alpha_{id},\upsilon_{pose}\right)$ and $\bar{T}\left(\beta_{exp},\upsilon_{pose},\alpha_{id}\right)$ represent the face pose and identity features obtained after the facial attributes pass through different encoder networks based on expression consistency.

Pose branch: The branching of the pose part aims to improve the alignment of 3D face landmarks by coupling the identity and pose attributes while preserving the consistency of the pose attribute $\bar{E}_{pose} = \mathcal{F}_{pose}\left(I;\upsilon_{pose}\right)$. Its joint attribute decomposition is expressed as:

$$\bar{E}_{pose,id} = \mathcal{F}_{pose,id}\left(\alpha_{id};\boldsymbol{I},\bar{E}_{pose}\right) \tag{17}$$

$$\bar{E}_{pose,id,exp} = \mathcal{F}_{pose,id,exp}\left(\beta_{exp};\boldsymbol{I},\bar{E}_{pose},\bar{E}_{pose,id}\right) \tag{18}$$

$$\bar{E}_{pose,exp} = \mathcal{F}_{pose,exp}\left(\beta_{exp};\boldsymbol{I},\bar{E}_{pose}\right) \tag{19}$$

$$\bar{E}_{pose,exp,id} = \mathcal{F}_{pose,exp,id}\left(\alpha_{id};\boldsymbol{I},\bar{E}_{pose},\bar{E}_{pose,exp}\right) \tag{20}$$

where $\bar{E}_{pose}$ is the pose attribute information obtained from the input image $\boldsymbol{I}$ after $\mathcal{E}_{pose}$ learning. $\bar{E}_{pose,id}$ and $\bar{E}_{pose,exp}$ are the identity and expression information obtained by $\mathcal{E}_{id}$ and $\mathcal{E}_{exp}$ encoders on the basis of the pose attribute $\bar{E}_{pose}$. $\bar{E}_{pose,id,exp}$ is the expression information obtained by $\mathcal{E}_{exp}$ encoder learning on the basis of $\bar{E}_{pose,id}$. By the same token, $\bar{E}_{pose,exp,id}$ is the identity information obtained by $\mathcal{E}_{id}$ encoder learning on the basis of $\bar{E}_{pose,exp}$. Therefore, the standard attribute representation value of identity and expression under the condition of the constant pose is :

$$\bar{T}_{(\upsilon_{pose},\alpha_{id},\beta_{exp})} = \bar{E}_{pose} \otimes \bar{E}_{pose,id} \otimes \bar{E}_{pose,id,exp} \tag{21}$$

$$\bar{T}_{(\upsilon_{pose},\beta_{exp},\alpha_{id})} = \bar{E}_{pose} \otimes \bar{E}_{pose,exp} \otimes \bar{E}_{pose,exp,id} \tag{22}$$

where $\bar{T}\left(\upsilon_{pose},\alpha_{id},\beta_{exp}\right)$ and $\bar{T}\left(\upsilon_{pose},\beta_{exp},\alpha_{id}\right)$ represent the face expression and identity features obtained after the facial attributes pass through different encoder networks based on pose consistency.

Fusion module: In order to reconstruct a more realistic and complete 3D human face. We extract the face information related to each attribute from the identity, expression, and pose branching networks, respectively, and use the fusion module to merge the face attributes to complete an accurate 3D face reconstruction. The purpose of doing so is to use this module to ensure the validity of face attribute decomposition. In addition, the face images that pass through the low-level network usually have higher resolution and contain more transparent information. Therefore, we introduce each branch's shallow face attribute information in the fusion module to better refine the face details. In which the face attribute features of the fusion module are represented as:

$$\begin{aligned} G_i = T_i\Big(\mathcal{R}\Big[&\bar{T}_{(\alpha_{id},\beta_{exp},\upsilon_{pose})}\Big), \bar{T}_{(\alpha_{id},\upsilon_{pose},\beta_{exp})}, \\ &\bar{T}_{(\beta_{exp},\alpha_{id},\upsilon_{pose})}, \bar{T}_{(\beta_{exp},\upsilon_{pose},\alpha_{id})}, \\ &\bar{T}_{(\upsilon_{pose},\alpha_{id},\beta_{exp})}, \bar{T}_{(\upsilon_{pose},\beta_{exp},\alpha_{id})}\Big]\Big) \end{aligned} \tag{23}$$

where $T_i(\cdot)$ denotes the feature information after the fusion of face attributes, $\mathcal{R}(\cdot)$ denotes the fusion of identity, expression, and pose attributes of the face, $Gi(\cdot)$is the feature output after fusion of face attributes.

D. *Objective Loss Function*

In the face attribute branching network, we introduce four learning objectives in model training. Among them, in learning face attribute parameters, we mainly constrain from three parts: identity, expression, and pose. To improve the network's effective learning of identity attributes, we optimize the parameters α_{id} by minimizing the vertex distance between the predicted face geometry and the ground truth 3D face as:

$$\begin{aligned}\mathcal{L}_{id} &= \left\|\left(S_{model} - \bar{S}_{model}\right)\right\| \\ &= \left\|\left(\Delta S_{id} - \Delta \bar{S}_{id}\right)\right\|^2 \\ &= \left\|\left(A_{id}\alpha_{id} - A_{id}\bar{\alpha}_{id}\right)\right\|^2\end{aligned} \tag{24}$$

where α_{id} denotes the predicted face identity parameter, and $\boldsymbol{\bar{\alpha}}_{id}$ is the ground truth parameter. A_{id} is the identity base of the 3D deformation model PCA. In addition, different α_{id} dimensions and singular values affect the face geometry differently. Therefore, among the identity branches, the constraint of $\mathcal{L}_{id}$ identity information can reduce the influence of essential dimensions, especially those with large singular values.

Similarly, to refine the expression details of the face, we use expression consistency loss to enhance expression preservation:

$$\begin{aligned}\mathcal{L}_{exp} &= \left\|\left(\Delta E_{exp} - \Delta \bar{E}_{exp}\right)\right\|^2 \\ &= \left\|\left(B_{exp}\beta_{exp} - B_{exp}\bar{\beta}_{exp}\right)\right\|^2\end{aligned} \tag{25}$$

where β_{exp} denotes the predicted face expression parameter, and $\bar{\beta}_{exp}$ is the expression ground truth parameter. B_{exp} is the expression base of the 3D deformation model PCA.

Since the pose of a face has limited degrees of freedom, to simplify the computation, we constrain the pose estimation by the loss of facial landmarks:

$$\begin{aligned}\mathcal{L}_{p} = \|&\left(f * P_r * R * \bar{S}_{model} + t\right) - \\ &\left(\bar{f} * P_r * \bar{R} * \bar{S}_{model} + \bar{t}\right)\|^2\end{aligned} \tag{26}$$

where $v_{pose} = [f, R, t]$ is the prediction of the pose parameters. $\bar{v}_{pose} = [\bar{f}, \bar{R}, t]$ is the ground truth of the pose parameter.Thus, we further improve the face landmark alignment accuracy under significant pose conditions by constraining the pose parameters. Although the $\mathcal{L}_{id}$, $\mathcal{L}_{exp}$, and $\mathcal{L}_{pose}$ loss functions have strong constraints on the identity, expression, and pose attributes of the three branches, the reconstructed 3D faces still lack the constraints of geometric contours. Therefore, we use the constraints of sparse 2D face landmark to improve the reconstructed geometric contour information as:

$$\begin{aligned}\mathcal{L}_{68} = \|&\left(f * P_r * R * S_{68} + t\right) - \\ &\left(\bar{f} * P_r * \bar{R} * \bar{S}_{68} + \bar{t}\right)\|^2\end{aligned} \tag{27}$$

The final loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \lambda_{id}\mathcal{L}_{id} + \lambda_{\mathrm{exp}}\mathcal{L}_{\mathrm{exp}} + \lambda_{p}\mathcal{L}_{p} + \lambda_{68}\mathcal{L}_{68} \tag{28}$$

where λ_{id}, λ_{exp},λ_{pose} and λ_{68} are the weights to balance these constraints.

IV. Experiments

In this section, we perform several experiments and ablation studies on DRTN using different settings on three extensively evaluated datasets, 300W-LP [22], AFLW2000-3D [22], and AFLW [38], to demonstrate the effectiveness of the method in 3D face reconstruction and dense face alignment. In addition, we further evaluate the generalization performance of the DRTN method on the HELEN [39], LS3D-W [41], and CelebA [40] datasets.

A. Implementation Details

We use the Pytorch [42] deep learning framework to train the DRTN model. To train our framework, the face regions are cropped according to the ground truth 3D facial landmarks and then scaled to 256 × 256 as the input to our network. Throughout the training process of the branching network, the increase in the number of layers and parameters of the network makes the neural network training slow and occasionally overfitting to the training set. This results in a deep cascade regression network that may not learn any information from the overfitted samples. In order to better balance the learning weights of each branch parameter, the loss weights of $\lambda_{id} = 1.0e^{-3}$, $\lambda_{exp} = 1.0e^{-4}$ and $\lambda_{pose} = 1.0e^{-4}$ are set for training after several experiments.

In addition, both training and testing experiments were conducted on a PC with NVIDIA GeForce and CUDA 11.2. The SGD solver set the minimum batch size and initial learning rate to 128 and 0.03, respectively. 687854 face images were available in our training set, of which 122450 natural face images and 565404 synthetic face images were available. The authentic face images are from the 300WLP [22] dataset, which is extended using various data enhancement algorithms. We performed a total of 70 batches of training. After 30, 40, and 60 batches, we reduced the learning rate to 0.02, 0.004, and 0.0008, respectively.

B. Datasets

The proposed DRTN is trained on 300W-LP [22] and evaluated on five popular datasets, AFLW [38], AFLW2000-3D [22], LS3D-W [41],and CelebA [40].

300W-LP: The 300W-LP dataset [23] is used to train our model, which is extended from 300W by standardizing multiple alignment datasets with 68 landmarks, including AFW [43], LFPW [44], HELEN [39], IBUG [45] and XM2VTS [56].

AFLW2000-3D: AFLW2000-3D [22] is constructed to evaluate 3D face alignment on challenging unconstrained images. This dataset contains the first 2000 images from AFLW and expands its annotations with fitted 3DMM parameters and 68 3D landmarks. We use this dataset to evaluate the performance of our method for the face alignment and face reconstruction task.

AFLW: AFLW [38] is a large-scale face dataset, including multiple poses and views, generally used to evaluate the effectiveness of facial landmark detection. The dataset has 25,993 face images with 21 landmarks annotated for each face, whereas landmarks are not annotated for faces in invisible regions. In addition, the dataset also includes face pose angle annotations obtained from the average 3D face reconstruction. Most of the face images in the AFLW dataset are color images; a few are grayscale images, of which 59% are female and 41% are male. This dataset is well suited for multi-angle multi-face detection, landmark localization, and head pose estimation,

TABLE I
THE $\mathrm{NME_{2D}}(\%)$ OF FACE ALIGNMENT RESULTS ON AFLW AND AFLW2000-3D

Method	AFLW Dataset (21 pts)					AFLW2000-3D Dataset (68 pts)				
	[0,30]	[30,60]	[60,90]	Mean	Std	[0,30]	[30,60]	[60,90]	Mean	Std
CDM [49]	8.150	13.020	16.170	12.440	4.040	-	-	-	-	-
RCPR [50]	5.430	6.580	11.530	7.850	3.240	4.260	5.960	13.180	7.800	4.740
ESR [51]	5.660	7.120	11.940	8.240	3.290	4.600	6.700	12.670	7.990	4.190
SDM [52]	4.750	5.550	9.340	6.550	2.450	3.670	4.940	9.760	6.120	3.210
DEFA [53]	-	-	-	-	-	4.500	5.560	7.330	5.803	1.169
3DDFA(CVPR2016) [22]	5.000	5.060	6.740	5.600	0.990	3.780	4.540	7.930	5.420	2.210
Yu et al.(ICCV2017) [54]	5.940	6.480	7.960	-	-	3.620	6.060	9.560	-	-
Nonlinear(CVPR2018) [31]	-	-	-	-	-	-	-	-	4.700	-
DAMDNet(ICCVW19) [13]	4.359	5.209	6.028	5.199	0.682	2.907	3.830	4.953	3.897	0.837
GSRN(MMM2021) [55]	4.253	5.144	5.816	5.073	0.638	2.842	3.789	4.804	3.812	0.801
MARN(ICPR2021) [14]	4.306	4.965	5.775	5.015	0.601	2.989	3.670	4.613	3.757	0.666
MFIRRN(ICASSP2021) [56]	4.321	5.051	5.958	5.110	0.670	**2.841**	**3.572**	4.561	3.658	0.705
DRTN(Ours)	**4.181**	**4.736**	**5.489**	**4.802**	**0.536**	2.861	3.592	**4.436**	**3.630**	**0.644**

which is an essential dataset inside the field of face landmark alignment.

LS3D-W: LS3D-W [41] is a large-scale face alignment annotation dataset created by the Computer Vision Laboratory at the University of Nottingham. The face images are from AFLW [38], 300-VW [47], 300-W [46], and FDDB [48]. Each face image in the dataset contains 68 annotated landmarks, containing a total of approximately 230,000 accurately labeled images of faces.

CelebA: CelebA [40] is a large-scale face attribute dataset containing over 200K images, each with 40 attribute annotations. The images in this dataset cover an extensive range of pose variations and complex backgrounds. There are 10177 identities and 202599 face images included in the CelebA dataset.

C. Evaluation

In terms of face alignment and reconstruction, we use $\mathrm{NME_{2d}}$ and $\mathrm{NME_{3d}}$ measurement methods to quantitatively evaluate the performance as:

$$\mathrm{NME_{2d}} = \frac{1}{N_k}\sum_{j}^{N_k}\left[\frac{1}{D_j}\sum_{i}^{N}\left(\left\|V_i^{2d}-\bar{V}_i^{2d}\right\|_2\right)\right] \tag{29}$$

$$\mathrm{NME_{3d}} = \frac{1}{N_k}\sum_{j}^{N_k}\left[\frac{1}{S_j}\sum_{i}^{N}\left(\left\|V_i^{3d}-\bar{V}_i^{3d}\right\|_2\right)\right] \tag{30}$$

where V_i^{2d} and $\bar{V}_i^{2d}$ denote the estimated 2D landmarks and the ground truth landmarks, respectively. we give the ground truth 3D vertices $\bar{V}_i^{3d}$ and the estimated vertices V_i^{3d} for N_k test images. D_j and S_j are the diagonal size of the face region in image space and 3D coordinate space, respectively. $\mathrm{NME_{2d}}$ evaluates the normalized 2D facial landmarks prediction error and $\mathrm{NME_{3d}}$ the evaluates the normalized 3D face geometry estimation accuracy. Due to the ambiguity of the weak perspective projection model, the reconstruction results of different methods have some ambiguity in the z-axis direction. We use rigid translation along the z-axis to align each result within the ground truth.

D. Analysis of 3D face alignment results

We quantitatively evaluate the face landmark alignment performance using normalized mean error $\mathrm{NME_{2D}}(\%)$ on the AFLW2000-3D [22] and AFLW [38] datasets. We divide the test set into three subsets in the test dataset based on the absolute yaw angle: $[0°, 30°]$, $[30°, 60°]$, and $[60°, 90°]$. Our DRTN was compared experimentally with the current state-of-the-art methods [13], [14], [22], [31], [49]–[56]. The results are shown in **Table 1** , with the best results in each category highlighted in bold, and the lower values of the results, the better. **Fig. 4** shows the corresponding CED curves, and our DTRN is only compared with the methods available for the codes in **Table 1** . The other methods are not comparable mainly because there is no relevant open-source code. Compared to benchmark methods [13], [14], [22], [53], [55], [56], the DTRN method has a lower normalized mean error on the AFLW [38] and AFLW-2000 [22] datasets. The experimental results show that the DTRN method can significantly improve the 3D face landmark alignment accuracy in the full pose range, and the face landmark alignment is also robust in large pose conditions.

Fig. 5 shows our method's 3D face landmark alignment results on the AFLW [38] AFLW2000-3D [22] datasets. The advantage of using 3DMM instead of other geometry representations is that normal mapping can associate the semantic facial landmarks that can be associated with the corresponding points in reconstructed geometry. The visualization results show that the DTRN method significantly outperforms 3DDFA [22],DAMDNet [13],GSRN [55], MARN [14], and MFRRN [56] methods for 3D face landmark alignment under large pose, extreme expression, and occlusion conditions, especially for eyes, mouth, and face contours. The qualitative results show that our DTRN method can significantly improve the network's learning of face attribute features in complex environments and thus improve the accuracy of face landmark alignment.

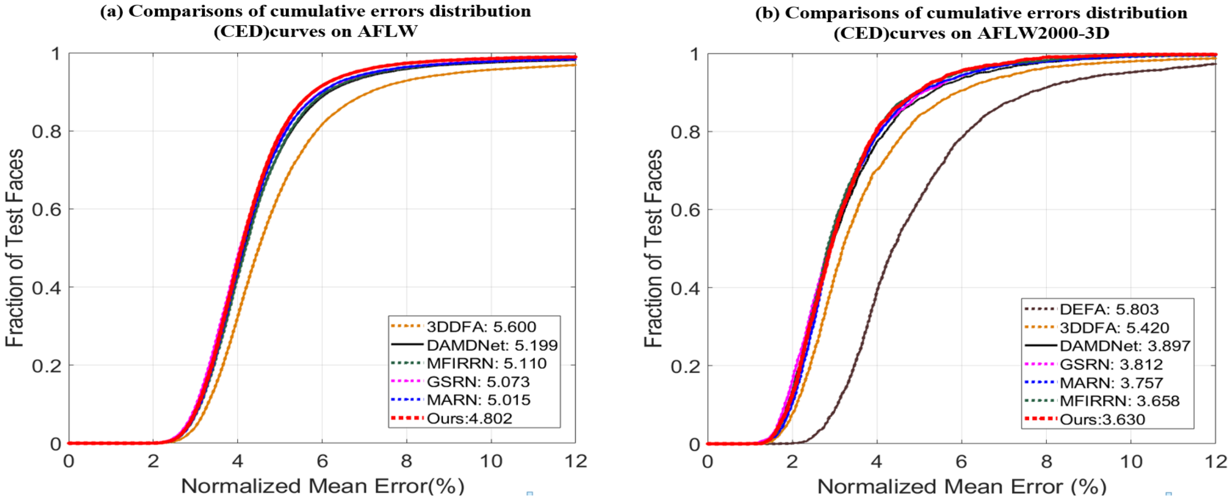


Fig. 4. The cumulative errors distribution (CED) curves on AFLW and AFLW2000-3D.

TABLE II

THE $\mathrm{NME}_{3d}(\%)$ OF FACE RECONSTRUCTION RESULTS ON AFLW2000-3D.

	3DDFA [22]	DAMDNet [13]	MFIRRN [56]	MARN [14]	GSRN [55]	DRTN(Ours)
$[0°, 30°]$	4.877	4.672	4.760	4.721	***4.543***	4.569
$[30°, 60°]$	6.086	5.619	5.488	5.535	5.368	***5.318***
$[60°, 90°]$	8.437	7.855	7.594	7.483	7.620	***7.120***
Mean	6.467	6.049	5.947	5.913	5.844	***5.669***
std	1.478	1.334	1.196	1.159	1.301	***1.070***

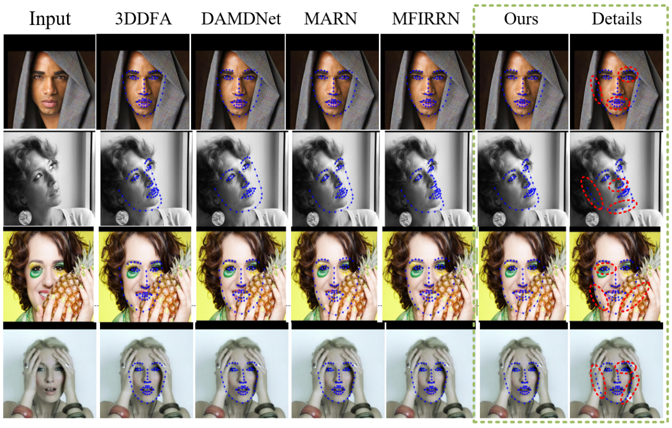


Fig. 5. Comparison of 3D facial landmark detection with 3DDFA [22], DAMDNet [13], MARN [14], MFRRN [56], and DRTN(Ours) on AFLW2000-3D. Best viewed on screen with zooming in.

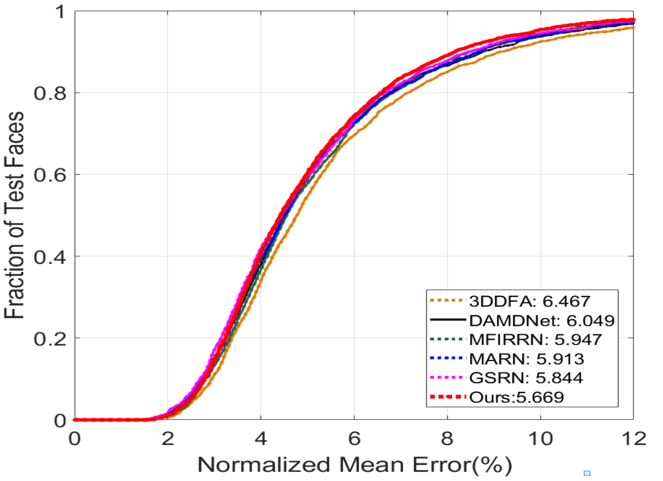


Fig. 6. The cumulative errors distribution (CED) curves on AFLW2000-3D.

E. Analysis of 3D face reconstruction results

The AFLW dataset is unsuitable for evaluating 3D face reconstruction because recovering 3D faces from annotated visible landmarks usually leads to ambiguity problems. To validate the effectiveness of our model in 3D face reconstruction, we compared the 3D normalized mean error $\mathrm{NME}_{3d}(\%)$ for the AFLW2000-3D dataset, as shown in **Table 2**. The first best result in each category is highlighted in bold, the lower is the better. The experimental results in **Table 2**, show that our DRTN method outperforms the state-of-the-art methods [13], [14], [22], [55], [56] for face reconstruction in both medium and large poses. Compared to the recent MARN [14], the $\mathrm{NME}_{3d}(\%)$ with our DRTN at offset angles $[0°, 30°]$, $[30°, 60°]$, and $[60°, 90°]$ is reduced by 3.2%, 3.9% and 4.9%, respectively. It can be seen that our method significantly improves the prediction accuracy at large offset angle $[60°, 90°]$, indicating that the model can reconstruct 3D faces accurately

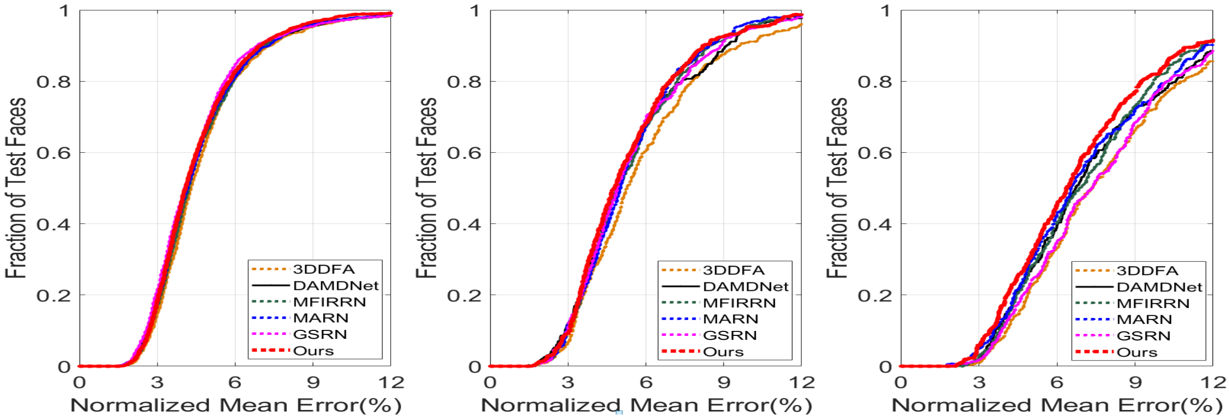


Fig. 7. The CED curve of the small, medium and large pose on AFLW2000-3D.

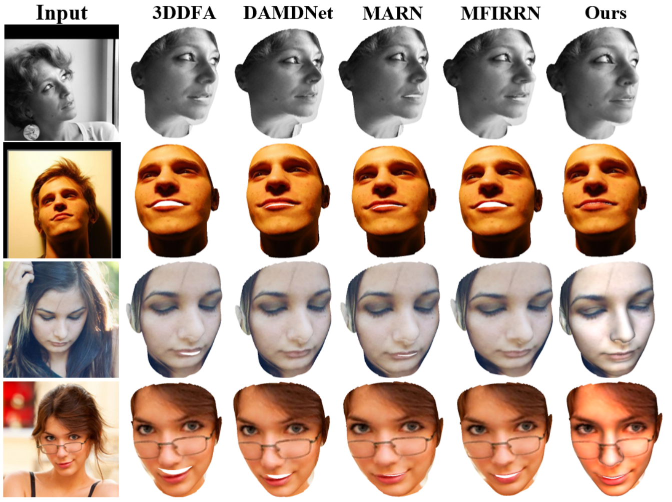


Fig. 8. Comparison of qualitative results of 3D face reconstruction on AFLW2000-3D. Images of the first column are the original face images. As can be seen from the image textures, the 3D face texture details reconstructed by our DRTN model are more natural and detailed.

even under unconstrained large pose conditions. Compared with GSRN [55], our DRTN can be reduced by 1.0% and 6.6% on $\mathrm{NME}_{3d}(\%)$ when the offset angles are medium pose $[30°, 60°]$ and large pose $[60°, 90°]$, respectively. This further validates the superior 3D reconstruction capability exhibited by our DRTN model in the case of complex occlusions. **Fig. 6** shows the corresponding CED for 3D face reconstruction, which instinctively proves that DRTN can achieve accurate 3D face reconstruction.

In order to prove the effectiveness of our method on face reconstruction in the large pose. In this experiment, we regard the whole AFLW2000-3D as the testing set and divide it into three subsets according to their absolute yaw angles:$[0°, 30°]$, $[30°, 60°]$, and $[60°, 90°]$ with 1312, 383, and 305 samples respectively. In **Fig. 7**, the CED curve results are shown from left to right for the small pose $[0°, 30°]$, the medium pose $[30°, 60°]$, and the large pose $[60°, 90°]$, respectively. Comparing the results of CED curves, our method is improved in small, medium, and large poses. The results fully validate the robustness of our DRTN algorithm for the reconstruction task in the large pose case. The visualization results are shown in **Fig. 8**. Our DRTN method reconstructs 3D faces with clear texture and natural expressions under large poses, extreme expressions, and partial occlusion. As shown in **Fig. 9**, the DRTN method is more geometrically accurate in terms of the reconstructed facial geometry compared to the current state-of-the-art methods, especially regarding local details of the eyes, mouth, and wrinkles. 3DDFA [22], DAMDNet [13], MARN [14], and MFIRRN [56] methods, because of the lack of face attribute correlation learning, do not reconstruct faces with fine expression details. Therefore, our DRTN approach greatly advantages high-fidelity 3D facial reconstruction under large poses, occlusion, and extreme expressions.

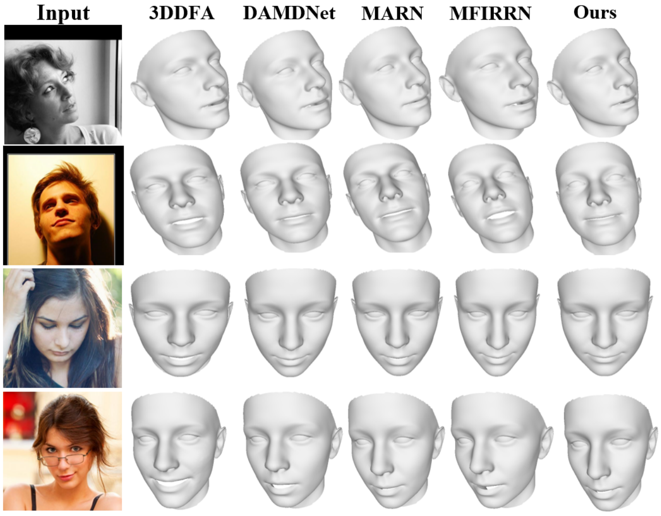


Fig. 9. Qualitative results of 3D face reconstruction on AFLW2000-3D. Best viewed on screen with zooming.

F. Ablation Experiments

To verify the effectiveness of each attribute branch module in DRTN, **Table 3** presents the ablation experiments on the

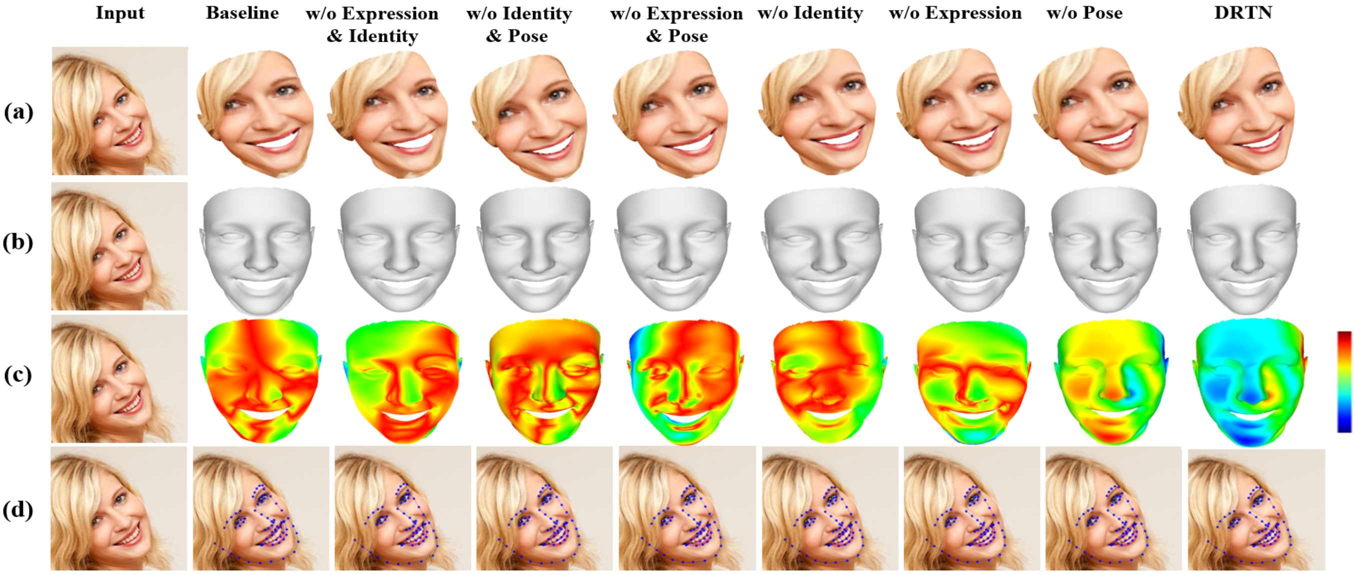


Fig. 10. Comparison of different network branching structures on the AFLW-2000 dataset. **(a)** and **(b)** are reconstructed high-fidelity 3D faces and facial geometries. **(c)** and **(d)** are error maps and face landmark alignment results.

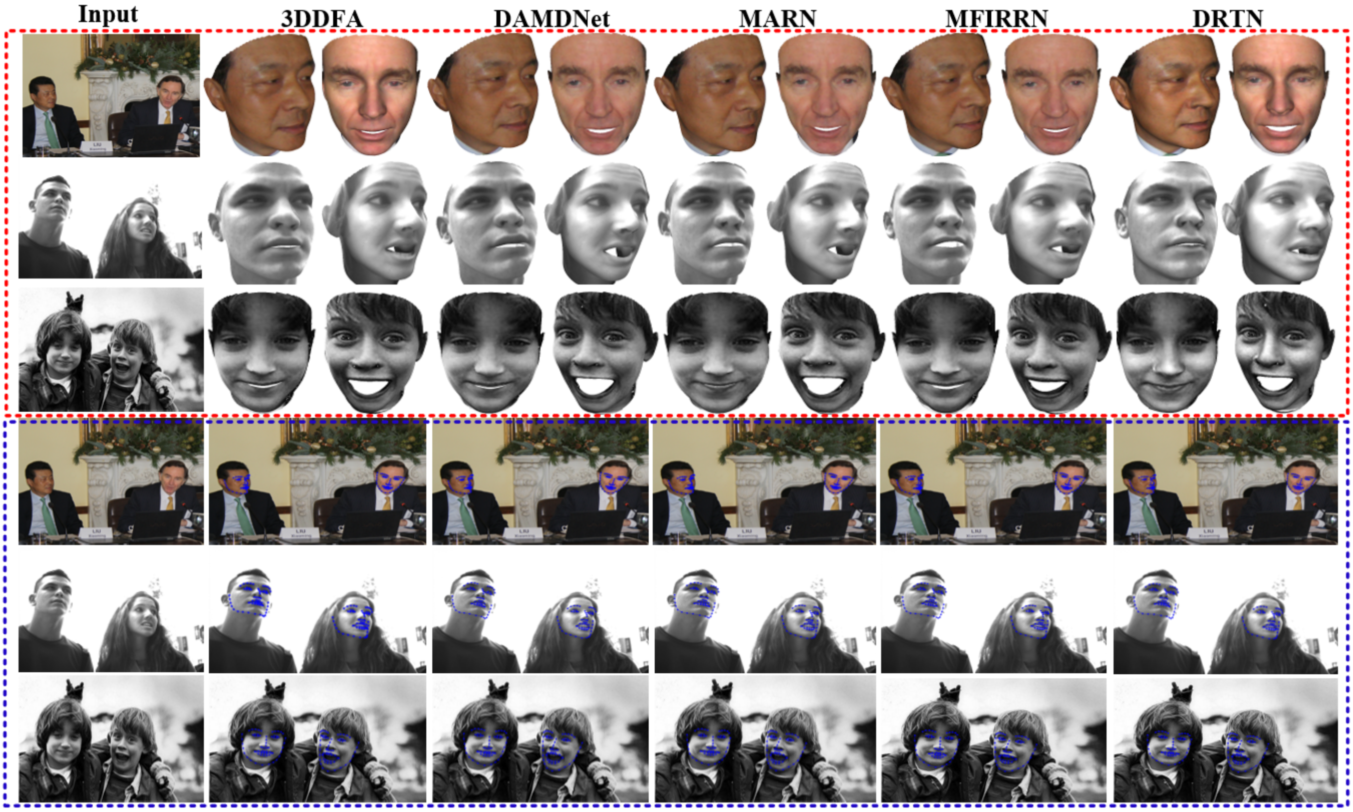


Fig. 11. Comparison of models of different methods on the LS3D-W dataset. In the red dashed box is the high-fidelity 3D face. The blue dashed box shows the 3D landmark alignment results.

TABLE III

ABLATION STUDY.THE NME(%) OF FACE ALIGNMENT AND RECONSTRUCTION RESULTS ON AFLW2000-3D FOR DIFFERENT NETWORK BRANCHING STRUCTURES.

Model	$\mathrm{NME}_{2\mathrm{D}}$	$\mathrm{NME}_{3\mathrm{D}}$
Baseline	4.450	6.563
DRTN w/o Identity and Expression Branchs	4.247	6.372
DRTN w/o Identity and Pose Branchs	4.077	6.132
DRTN w/o Expression and Pose Branchs	3.899	6.125
DRTN w/o Identity Branch	3.828	5.967
DRTN w/o Expression Branch	3.772	5.769
DRTN w/o Pose Branch	3.697	5.674
DRTN	**3.630**	**5.669**

AFLW-2000 dataset, where the normalized mean errors of face landmark alignment and reconstruction are $\mathrm{NME}_{2\mathrm{D}}$ and $\mathrm{NME}_{3\mathrm{D}}$, respectively. We can see by comparing the first to fourth rows of **Table 3** that the network with the face pose, expression, and identity attribute branches has lower reconstruction and alignment errors than the baseline model without any face attribute branches. The reconstruction and alignment errors are lower than those of the baseline model without face attribute branches. The experimental results show that the disentangled representation of face attributes can improve the model's attribute learning ability and robustness. In addition, from the second to the fourth row of **Table 3**, it can be found that the identity attribute model outperforms the expression and pose attribute models in the single-branch face attribute network because the number of identity attribute parameters in the labels of the training dataset is more than the expression and pose parameters. Therefore, when the label has more labeling information in a single-branch face attribute network, it is more beneficial for the network to learn the attribute parameters. Similarly, we can find from the fifth to seventh rows of **Table 3** that the network model using two-branch face attributes has lower face alignment and reconstruction errors than the single-branch face attribute network. The results further indicate that using face attribute disentangled representation can improve the network's learning of face attribute features. Finally, the error results in the table show that the DRTN method we designed shows good attribute learning ability in face alignment and reconstruction. The main reason is that the information between face attributes is correlated with each other and does not exist singularly. We can enrich the details of face reconstruction with the correlation of face attributes.

In addition, it further demonstrates the role played by face attribute branching networks in face landmark alignment and reconstruction. We visualize the face landmark alignment and reconstruction results for each branch network, and the results are shown in **Fig.10**. From the reconstructed high-fidelity 3D face and geometry results, we can see that the 3D face reconstructed by our DRTN method has a natural expression and fits the shape of the original image and the geometry of the face edges and the contours of the five features are clear. We visualize the reconstructed 3D faces of each model in the form of heat maps. The heat map shows that the 3D faces reconstructed by our method are more realistic and have fewer errors. Regarding 3D face landmark alignment, the DRTN method also has higher landmark alignment accuracy than other face attribute branching models in large poses. Because other face attribute branching networks lack mutual learning between attributes in the learning process of face attribute parameters, although the network learns independent face attributes well, the accuracy of reconstruction and alignment in complex and diverse situations could be better because the 3D face is not a linear model. In contrast, the DRTN model incorporates identity, expression, and pose information and therefore shows good stability in reconstructing unconstrained scenes.

G. A visualization experiment performed on LS3D-W

This section compares the qualitative results of our DRTN method with the most appropriate state-of-the-art methods in the LS3D-W [41]. **Fig .11** shows the 3D face reconstruction and landmark alignment results in complex scenes with insufficient illumination and a large pose of the face. As can be seen from the red dashed box in the figure, the DRTN model has higher reconstruction accuracy than 3DDFA [22], DAMDNet [13], MARN [14], and MFIRRN [56] models in high-fidelity 3D face reconstruction, especially in terms of facial features contour, which validates that our model can better capture local details and geometric contour information. The LS3D-W [41] is collected in a high-level unconstrained complex environment with different illumination and face offset angles. The blue dashed box in the figure shows the face landmark alignment results. The DRTN model can still show good alignment results under self-obscuring faces and low-light conditions. Therefore, the excellent experimental results verify the powerful 3D reconstruction capability and landmark alignment accuracy of the face attribute disentangling representation method.

H. Qualitative Results on CelebA and Casual Photos

To further validate the generalization ability of our DRTN on other datasets, we used casual photos and the CelebA [40] for evaluation. As shown in **Fig.12**, the face images in the green dashed box are the experimental results of the CelebA [40]. The face images in the red dashed box are from some random photos of life and photos of anime characters. Our DRTN accomplishes accurate landmark alignment and fine 3D face reconstruction on both CelebA [40] and casual photos, reflecting its good generalization ability. Because the DRTN method allows the network to learn the correlation between identity, expression, and pose attributes and fuse the underlying information between face attributes. Therefore, accurate high-fidelity 3D faces and 3D face landmark alignment can be reconstructed.

CONCLUSION

In this paper, we propose a disentangled representation transformer network capable of recovering detailed 3D faces in an unconstrained environment and performing face dense

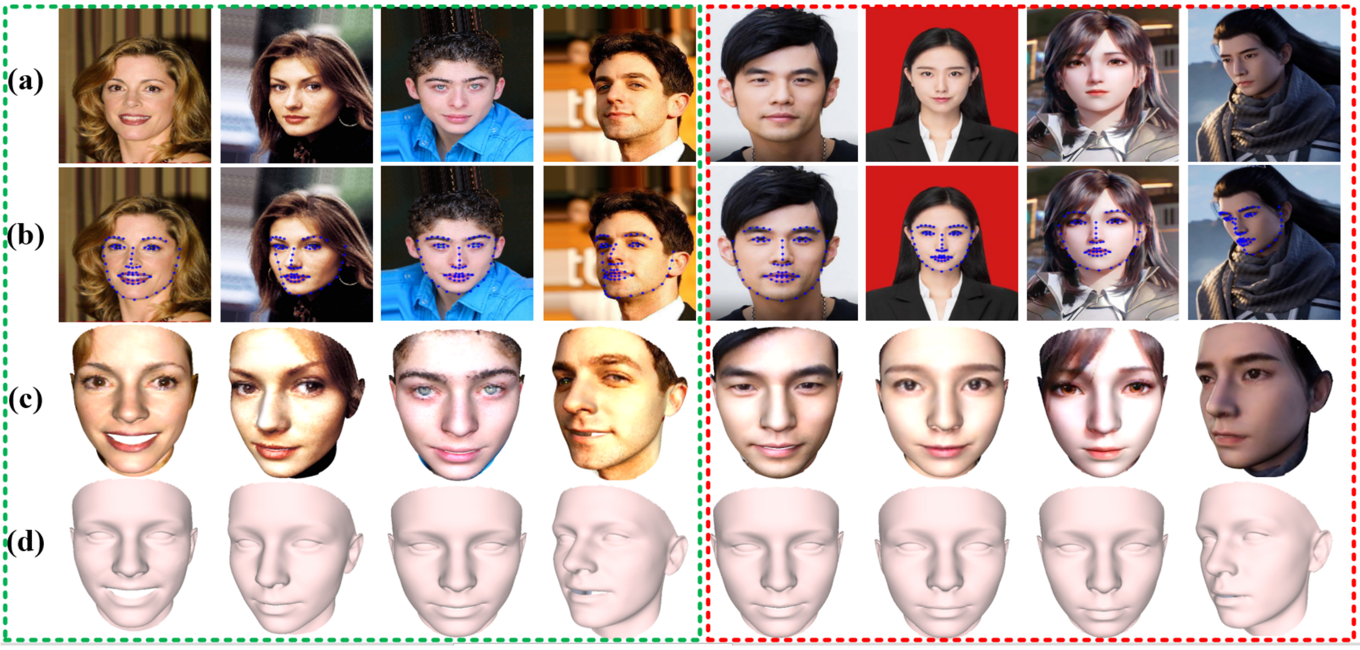


Fig. 12. Qualitative results on casual photos and CelebA. **(a)** Moreover, **(b)** are the original map and 3D landmark alignment results of the input. **(c)** Moreover, **(d)** is the high-fidelity 3D face and facial geometry.

alignment more accurately. Our DRTN method enhances the network's learning of potential information about face attributes and addresses the effects of facial expression, head pose, and partial occlusion on reconstruction and landmark alignment. Quantitative results show that the proposed DRTN model is more accurate than state-of-the-art face dense alignment and 3D reconstruction methods. In addition, extensive qualitative experiments show that DRTN can successfully reconstruct high-fidelity 3D faces from 2D face images with rich details and strong generalization ability. In future work, we will further investigate the proposed method in the context of 3D face reconstruction, such as face reconstruction of videos and cartoon character reconstruction, to further evaluate the generalization ability of the proposed method.

References

[1] Liu, F., Zhu, R., Zeng, D., Zhao, Q., Liu, X. (2018). Disentangling features in 3D face shapes for joint face reconstruction and recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5216-5225).

[2] Hu, J., Lu, J., Tan, Y. P. (2014). Discriminative deep metric learning for face verification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1875-1882).

[3] Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722.

[4] Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K. (2016). Real-time facial animation with image-based dynamic avatars. ACM Transactions on Graphics, 35(4).

[5] Amberg, B., Romdhani, S., Vetter, T. (2007, June). Optimal step nonrigid ICP algorithms for surface registration. In 2007 IEEE conference on computer vision and pattern recognition (pp. 1-8). IEEE.

[6] Kemelmacher-Shlizerman, I., Basri, R. (2010). 3D face reconstruction from a single image using a single reference face shape. IEEE transactions on pattern analysis and machine intelligence, 33(2), 394-405.

[7] Li, C., Zhou, K., Lin, S. (2014, September). Intrinsic face image decomposition with human face priors. In European conference on computer vision (pp. 218-233). Springer, Cham.

[8] Cao, X., Chen, Z., Chen, A., Chen, X., Li, S., Yu, J. (2018). Sparse photometric 3D face reconstruction guided by morphable models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4635-4644).

[9] Tuan Tran, A., Hassner, T., Masi, I., Medioni, G. (2017). Regressing robust and discriminative 3D morphable models with a very deep neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5163-5172).

[10] Jourabloo, A., Liu, X. (2016). Large-pose face alignment via CNN-based dense 3D model fitting. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4188-4196).

[11] Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In Proceedings of the European conference on computer vision (ECCV) (pp. 534-551).

[12] Jackson, A. S., Bulat, A., Argyriou, V., Tzimiropoulos, G. (2017). Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In Proceedings of the IEEE international conference on computer vision (pp.1031-1039).

[13] Jiang, Z. H., Wu, Q., Chen, K., Zhang, J. (2019). Disentangled representation learning for 3d face shape. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11957-11966).

[14] Li, X., Wu, S. (2021, January). Multi-attribute regression network for face reconstruction. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 7226-7233). IEEE.

[15] Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham, J. (1995). Active shape models-their training and application. Computer vision and image understanding, 61(1), 38-59.

[16] Cootes, T. F., Edwards, G. J., Taylor, C. J. (2001). Active appearance models. IEEE Transactions on pattern analysis and machine intelligence, 23(6), 681-685.

[17] Cootes, T. F., Ionita, M. C., Lindner, C., Sauer, P. (2012, October). Robust and accurate shape model fitting using random forest regression voting. In European conference on computer vision (pp. 278-291). Springer, Berlin, Heidelberg.

[18] Jourabloo, A., Liu, X. (2015). Pose-invariant 3D face alignment. In Proceedings of the IEEE international conference on computer vision (pp. 3694-3702).

[19] Bulat, A., Tzimiropoulos, G. (2016, October). Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw)

challenge. In European Conference on Computer Vision (pp. 616-624). Springer, Cham.

[20] Jeni, L. A., Cohn, J. F., Kanade, T. (2015, May). Dense 3D face alignment from 2D videos in real-time. In 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG) (Vol. 1, pp. 1-8). IEEE.

[21] Ning, X., Duan, P., Li, W., Zhang, S. (2020). Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer. IEEE Signal Processing Letters, 27, 1944-1948.

[22] Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 146-155).

[23] Liu, F., Zeng, D., Zhao, Q., Liu, X. (2016, October). Joint face alignment and 3D face reconstruction. In European Conference on Computer Vision (pp. 545-560). Springer, Cham.

[24] Jiang, L., Wu, X. J., Kittler, J. (2019). Dual attention MobDenseNet (DAMDNet) for robust 3D face alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).

[25] Ye, H., Lv, L., Liu, Y., Zhou, Y. (2016). Evaluation of the Accuracy, Reliability, and Reproducibility of Two Different 3D Face-Scanning Systems. The International Journal of Prosthodontics, 29(3), 213-218.

[26] De Wansa Wickramaratne, V. K., Ryazanov, V. V., Vinogradov, A. P. (2009). Analysis of a 3D face-scanning system by active triangulation. Pattern Recognition and Image Analysis, 19(1), 78-83.

[27] Blanz, V., Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques (pp. 187-194).

[28] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T. (2009, September). A 3D face model for pose and illumination invariant face recognition. In 2009 sixth IEEE international conference on advanced video and signal based surveillance (pp. 296-301). Ieee.

[29] Richardson, E., Sela, M., Or-El, R., Kimmel, R. (2017). Learning detailed face reconstruction from a single image. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1259-1268).

[30] Dou, P., Shah, S. K., Kakadiaris, I. A. (2017). End-to-end 3D face reconstruction with deep neural networks. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5908-5917).

[31] Tran, L., Liu, X. (2018). Nonlinear 3d face morphable model. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7346-7355).

[32] Lee, G. H., Lee, S. W. (2020). Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6100-6109).

[33] Wu, S., Rupprecht, C., Vedaldi, A. (2020). Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1-10).

[34] Browatzki, B., Wallraven, C. (2020). 3fabrec: Fast few-shot face alignment by reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6110-6120).

[35] Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In Proceedings of the European conference on computer vision (ECCV) (pp. 534-551).

[36] Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics, 20(3), 413-425.

[37] Xia, X., Li, J., Wu, J., Wang, X., Wang, M., Xiao, X., ... Wang, R. (2022). TRT-ViT: TensorRT-oriented Vision Transformer. arXiv preprint arXiv:2205.09579.

[38] Koestinger, M., Wohlhart, P., Roth, P. M., Bischof, H. (2011, November). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In 2011 IEEE international conference on computer vision workshops (ICCV workshops) (pp. 2144-2151). IEEE.

[39] Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In Proceedings of the IEEE international conference on computer vision workshops (pp. 386-391).

[40] Liu, Z., Luo, P., Wang, X., Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[41] Bulat, A., Tzimiropoulos, G. (2017). How far are we from solving the 2d 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision (pp. 1021-1030).

[42] Paszke, A., Gross, S., Chintala, S., Chanan, G. (2017). Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration, 6(3), 67.

[43] Zhu, X., Ramanan, D. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In 2012 IEEE conference on computer vision and pattern recognition (pp. 2879-2886). IEEE.

[44] Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. IEEE transactions on pattern analysis and machine intelligence, 35(12), 2930-2940.

[45] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the IEEE international conference on computer vision workshops (pp. 397-403).

[46] 300 Faces in-the-Wild Challenge, accessed on Jul.2013.[Online]. Available:http://ibug.doc.ic.ac.uk/resources/300-W/.

[47] Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In Proceedings of the IEEE international conference on computer vision workshops (pp. 50-58).

[48] Jain, V., Learned-Miller, E. (2010). Fddb: A benchmark for face detection in unconstrained settings (Vol. 2, No. 6). UMass Amherst technical report.

[49] Yu, X., Huang, J., Zhang, S., Metaxas, D. N. (2015). Face landmark fitting via optimized part mixtures and cascaded deformable model. IEEE transactions on pattern analysis and machine intelligence, 38(11), 2212-2226.

[50] Burgos-Artizzu, X. P., Perona, P., Dollár, P. (2013). Robust face landmark estimation under occlusion. In Proceedings of the IEEE international conference on computer vision (pp. 1513-1520).

[51] Cao, X., Wei, Y., Wen, F., Sun, J. (2014). Face alignment by explicit shape regression. International journal of computer vision, 107(2), 177-190.

[52] Yan, J., Lei, Z., Yi, D., Li, S. (2013). Learn to combine multiple hypotheses for accurate face alignment. In Proceedings of the IEEE international conference on computer vision workshops (pp. 392-396).

[53] Liu, Y., Jourabloo, A., Ren, W., Liu, X. (2017). Dense face alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 1619-1628).

[54] Yu, R., Saito, S., Li, H., Ceylan, D., Li, H. (2017). Learning dense facial correspondences in unconstrained images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4723-4732).

[55] Wang, X., Li, X., Wu, S. (2021, June). Graph structure reasoning network for face alignment and reconstruction. In International Conference on Multimedia Modeling (pp. 493-505). Springer, Cham.

[56] Li, L., Li, X., Wu, K., Lin, K., Wu, S. (2021, June). Multi-granularity feature interaction and relation reasoning for 3d dense alignment and face reconstruction. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4265-4269). IEEE.